A Deep Learning Framework for Automated Pap Smear Analysis in Cervical Cancer Screening

Arjun Garg, Shaurya Bisht, Gautham Agilan
Louis Vaickus MD PhD, Joshua Levy PhD
Emerging Diagnostic and Investigative Technologies, Department of Pathology, Dartmouth Hitchcock Medical Center

INTRODUCTION

We developed a pipeline for Pap smear whole slide image analysis. The process applies Macenko stain normalization to correct staining variability, followed by patch extraction and feature generation using a pretrained ResNet-50 model (2048-D vectors). These features are grouped using k-means clustering to identify morphologically similar patches, and a binary classifier distinguishes squamous from non-squamous regions. The approach was tested on Pap smear WSIs and performed for automated cytology analysis.

RESEARCH CONTEXT

Cervical cancer is a major global health challenge, causing over 300,000 deaths annually despite being highly preventable through early detection. Pap smear cytology remains the most widely used screening method, but its interpretation is highly subjective, time-consuming, and prone to human error. Variability between pathologists and differences in laboratory staining protocols can lead to inconsistent diagnoses and missed abnormalities.

Previous research has introduced digital pathology and machine learning approaches to assist in Pap smear analysis. These methods have demonstrated potential for automating abnormal cell detection and reducing manual workload. However, many earlier systems were limited to small regions of interest or focused primarily on isolated cells, failing to capture the complexity of whole-slide images. Moreover, stain variability and lack of standardization have remained persistent challenges, reducing the generalizability of such models across different clinical settings.

This research aims to build on these prior efforts by addressing these limitations and moving toward a more reliable, reproducible, and scalable approach to cervical cancer screening using computational methods.

METHODS

We applied our AI-based pipeline to a dataset of approximately 5,000 digitized Pap smear whole slide images (WSIs). All data came from partner hospitals, pathology labs, and screening programs using IRB-approved, clinical records. To prepare the WSIs for analysis, we performed several preprocessing steps:

- Tiling slides into 256×256 pixel patches to handle large images efficiently
- Macenko stain normalization to standardize color distributions shown in Fig. 1.
- We also conducted quality control to ensure consistent annotations across samples.

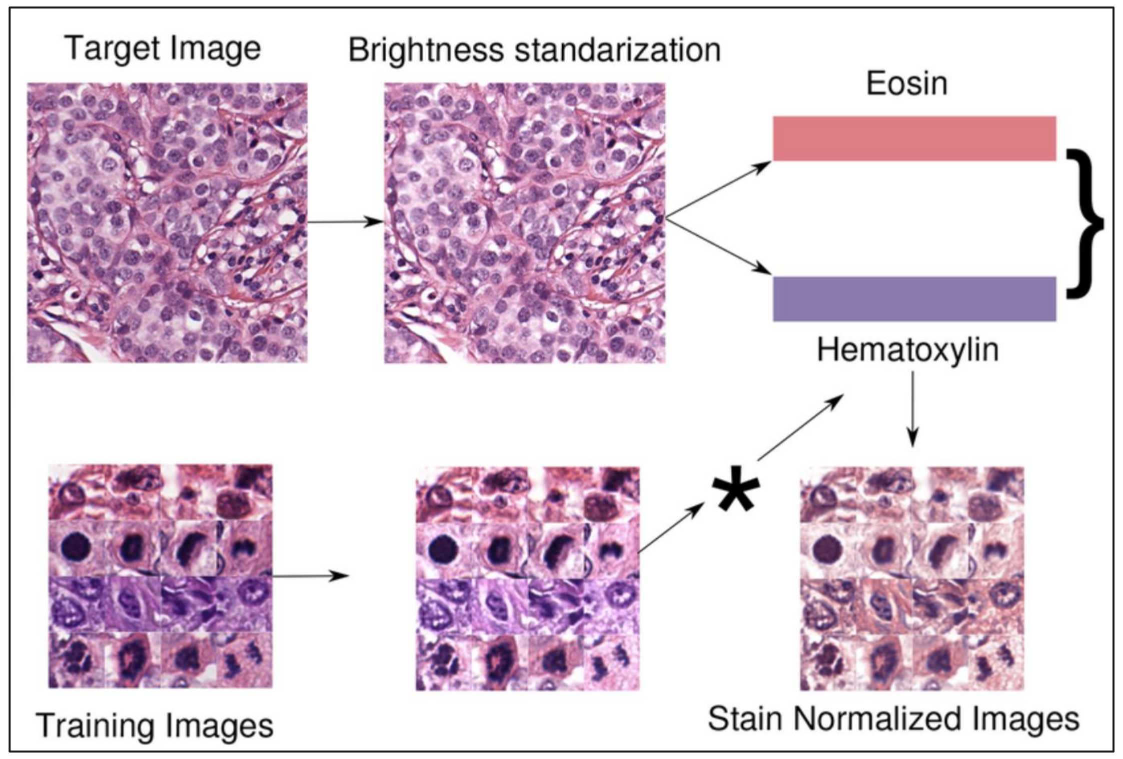


Figure 1. Diagram of pipeline for Macenko stain normalization

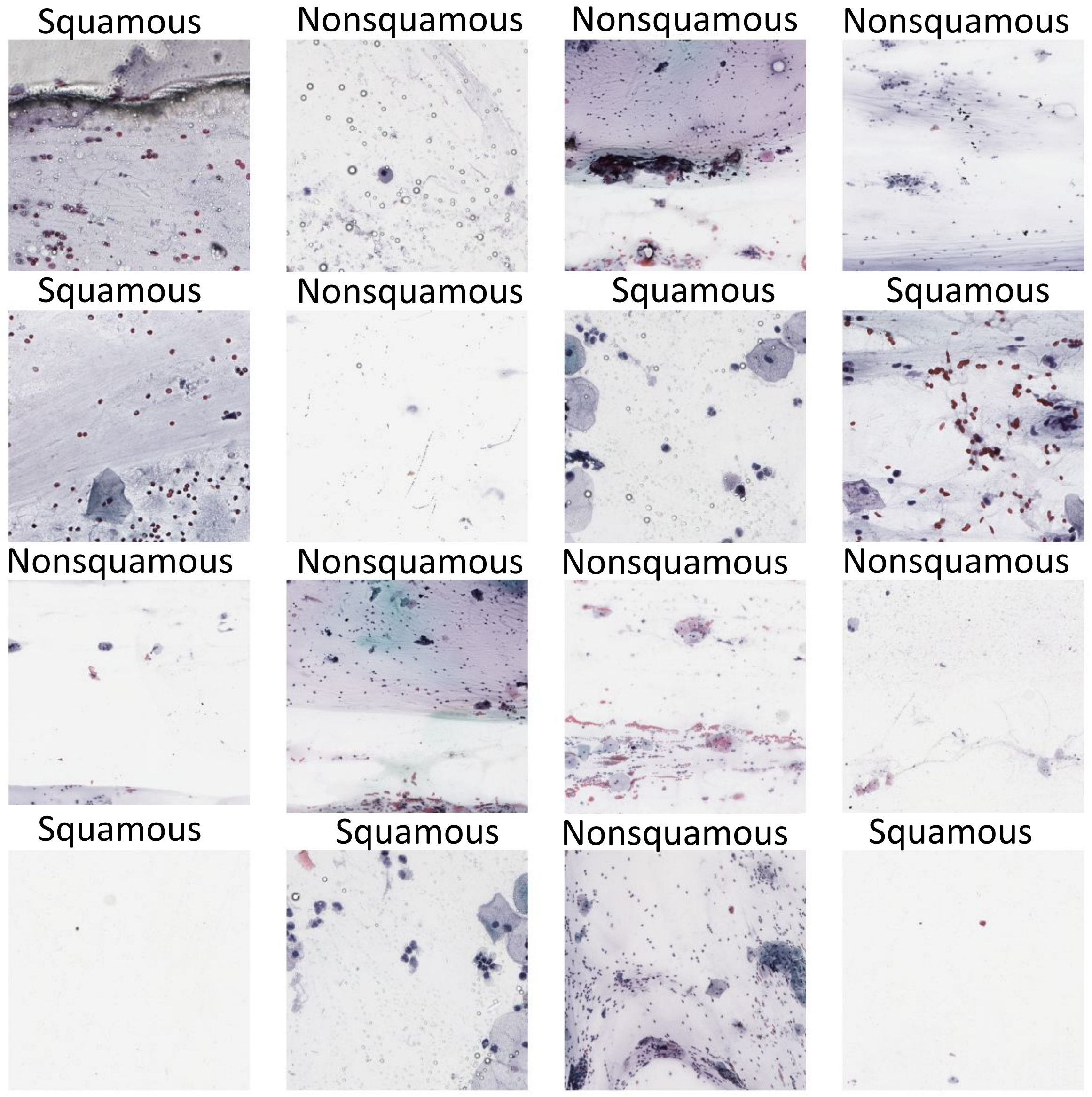
Next, each patch was processed using a pretrained ResNet-50 model with the classification head removed.

- Generated feature vectors for each patch
- Summarizes important patterns, like cell shapes and nuclei.
- Feature vectors are then used to group similar patches and classify them as squamous or non-squamous.

To explore structural similarities, we applied k-means clustering to the feature vectors. This grouped morphologically similar patches and facilitated unsupervised training. Finally, a binary classifier was trained on the extracted features to distinguish squamous from non-squamous regions. Training utilized cross-entropy loss, the Adam optimizer, and class balancing to handle label imbalances.

RESULTS

Our ResNet-50 model, fine-tuned on ~5,000 image patches, achieved strong performance in distinguishing squamous from non-squamous regions (Accuracy: 0.87, Precision: 0.86, Recall: 0.88, AUC: 0.92). Clustering with UMAP and K-means successfully grouped patches into meaningful categories, highlighting squamous subtypes and filtering out artifacts.



CONCLUSION

Our project demonstrates the potential of deep learning to streamline cervical cancer screening by automating patch-level classification and clustering of Pap smear WSIs. The pipeline we created reduces variability, supports early detection, and alleviates the workload on pathologists. Future steps include expanding the pipeline to incorporate detection and segmentation, validating on external datasets, and collaborating with pathologists to ensure real-world applicability.